Recherches sur Diderot et sur l'Encyclopédie

31-32 | Avril 2002
L'Encyclopédie en ses nouveaux atours électroniques: vices et vertus du virtuel

Le site ATILF

The ATILF Website

Zina Tucsnak

Édition électronique
URL : http://journals.openedition.org/rde/16
DOI : 10.4000/rde.16
ISSN : 1955-2416

Éditeur
Société Diderot

Édition imprimée
Date de publication : 15 avril 2002
Pagination : 27-30
ISSN : 0769-0886

Référence électronique
Zina Tucsnak, « Le site ATILF », *Recherches sur Diderot et sur l'Encyclopédie* [En ligne], 31-32 | Avril 2002, document 2, mis en ligne le 13 juin 2006, consulté le 02 mai 2019. URL : http://journals.openedition.org/rde/16 ; DOI : 10.4000/rde.16

Zina TUCSNAK

Le site ATILF

1. Introduction

Dans les dernières années, notre laboratoire et ARTFL (Université de Chicago) ont mis au point, chacun de son côté, un nombre important de ressources linguistiques informatisées. Au début de l'année 1999, la direction du laboratoire a décidé de coordonner ces efforts et de rassembler un grand nombre de ces outils. On peut citer : l'implémentation de l'*Encyclopédie* de Diderot, trois éditions du *Dictionnaire de l'Académie*, ainsi que les dictionnaires anciens de Nicot et de Bayle.

Instrument de recherche à la disposition de la communauté scientifique, l'*Encyclopédie* de Diderot et D'Alembert est une ressource linguistique accessible sur la toile française depuis mars 2000 en tant que produit du laboratoire ATILF (Analyses et traitements informatiques du lexique français), ancien INaLF.

Ressource linguistique soumise à un abonnement, l'*Encyclopédie* de Diderot et D'Alembert est accessible depuis le début de l'année dernière au CNRS, à l'ATILF à l'adresse : http://encyclopedie.inalf.fr/. La version de l'ATILF de l'*Encyclopédie* de Diderot et D'Alembert a pour but de transformer l'informatique en un vrai outil de travail convivial et puissant.

Cette communication cherchera à prouver sur des exemples (dans le cadre d'une démonstration) les différences, les limites, les atouts de chacun de ces outils qui portent sur l'*Encyclopédie* de Diderot et D'Alembert et les dictionnaires d'autrefois qui, dès leur parution, étaient une vraie toile...

*2. L'*Encyclopédie *de Diderot et D'Alembert*

La première version électronique de cet ouvrage de référence a été réalisée par l'équipe du Projet ARTFL, à l'Université de Chicago. Après la mise en place physique d'un ordinateur sous Linux et l'installation des logiciels nécessaires, on a installé l'implémentation de l'*Encyclopédie*

Diderot sous Philologic. Philologic est un moteur de recherche pour les bases textuelles. C'est un outil qui englobe des scripts Perl qui communiquent avec les butineurs www via la traditionnelle interface CGI et des programmes écrits en langage C.

Une première étape a été la réorganisation du site américain ainsi que la francisation des pages. Le site nancéien améliore la présentation des textes et facilite la recherche. Dans une seconde étape on a amélioré l'accès aux planches. Finalement, la modification des scripts pour une machine Linux adéquate fait que le site http://encyclopedie.inalf.fr/ est très rapide et compte de plus en plus de visiteurs.

Cette base de données est formée de plusieurs bases de données coordonnées : une base de données objet, une base de données pour l'élaborations des mots, un index de concordance des mots (qui reconnaît les objets textuels) et un gestionnaire d'objet. Le mécanisme des interrogations plein texte est identique à celui par mot vedette à l'exception du traitement des méta-données. La recherche plein texte a pour but d'exploiter les 15 éléments de la spécification Dublin Core. L'extraction du contenu des en-têtes Dublin Core se fait conformément aux spécifications ATE (ARTFL Text Encoding). Les spécifications ATE combinent le Dublin Core avec le HTML de base. Le codage permet quelques extensions : identificateurs de page, balises spécifiques pour les fins de phrase, balises pour les noms propres. ATE est un système-dépendant et il a une construction « bottom-up ». Le système ignore l'encodage SGML qui n'est pas explicitement traité par programme. Un des avantages est l'utilisation des éditeurs HTML, etc., pour la modification des bases. La base est construite par le moteur de recherche et le gestionnaire d'objets à partir d'une liste de fichiers. Des scripts Perl définissent les formats de sortie, les en-têtes généraux, les points de navigation et les renvois.

Le système permet toute recherche bibliographique : par mot-vedette, par auteur, etc., ainsi qu'une recherche par mot adjacent (pattern matching). La recherche plein texte permet les expressions régulières et les opérateurs logiques.

Pour l'*Encyclopédie*, les textes sont traités comme une hiérarchie d'objets textuels. A part les volumes, il y a les articles (72 000), les sous-articles, les paragraphes, les phrases et les mots (environ 21 millions), tout cela en parallèle avec d'autres structures : images, titres de pages, planches (3000), etc. La recherche comporte 4 parties :

- Définition du corpus (sélection d'objet)
- Mot (co-occurrence, pattern matching)
- Recherche dans les index
- Extraction du contexte et formatage final du format de sortie (conversion des balises SGML dans du HTML, résolution des références croisées, etc.).

Cette base fait partie des produits ATILF (avec la base FRANTEXT) soumis à un abonnement de 2 000 FF.

La version-microfiche publiée par IDC (Pays-Bas) qui a été utilisée pour le texte de l'*Encyclopédie* est, avec la garantie de Richard N. Schwab, une bonne reproduction de la première édition de l'*Encyclopédie* : l'édition de Paris. La version microprint a toutes les qualités d'une bonne première édition telle qu'elle a été définie par Richard N. Schwab (voir http://encyclopedie.inalf.fr/caveat.html). La saisie a été effectuée en Chine et il y en a encore beaucoup d'erreurs de saisie (1 erreur pour 15 000 caractères).

2.1 Quelques exemples

Une recherche plein texte sur « spinoza » ne donne que deux occurrences et des mots comme agnostique, rationalité, sensualisme, spinosiste, spinosisme n'y figurent pas. Une recherche croisée sur « théocratie » et « oeconomie » mène à Boulanger, ingénieur des Ponts et Chaussées, ami de Diderot[1].

2.2 Comparaison avec d'autres éditions

Une version « grand public » de *l'Encyclopédie* est éditée sous la forme de quatre cédéroms par les édition Redon. L' édition utilisée est celle de Genève. L'ambition de la société Redon : une faute pour 40 000 caractères. Une recherche sur « Jaucourt » donne 26 occurrences tandis qu'une recherche plein texte sur le site « http://encyclopedie.inalf.fr » donne 1 067 occurrences. L'avantage de cette édition est que les utilisateurs de tous les jours ont l'*Encyclopédie* chez eux pour un prix modique.

3. Développements futurs

- Le remplacement des scripts Perl qui composent le moteur de recherche par le standard SQL (sous sa forme Linux MySQL) qui est mieux adapté aux très grandes bases de données, rapide et fiable.
- Concevoir et réaliser une interface unique qui permettra des interrogations simultanées sur l'ensemble du corpus des dictionnaires d'autrefois.
- La mise en place d'un Forum et d'un espace FAQ (Foire aux Questions).
- Une stratégie à long terme : après une correction des erreurs liées au « méta-texte », une correction plus en profondeur, avec les Éditions Champion.

1. Tous ces exemples, qui ont été soulevés au cours du colloque, sont discutés dans les communications suivantes.

4. Bibliographie

Mark Olsen, *Text Theory and Coding Practice : Assessing the TEL*, 1996 Joint Annual Conference of the Association for Computers and the Humanities and Association for Literary and Linguistic Computing, Bergen, Norway, June 1996.

Leonid Andreev, Mark Olsen, « Conception de systèmes Hypermedia à grande échelle pour les sciences humaines : présentation de Philologic, le logiciel d'ARTFL », 67e Congrès de 1'ACFAS (l'Association canadienne-française pour l'avancement des sciences), May 11, 1999, University of Ottawa.

Zina TUCSNAK
ATILF-CNRS